Towards multilingual articulatory feature recognition with Support Vector Machines

Jan Macek, Anja Geumann and Julie Carson-Berndsen
School of Computer Science and Informatics, University College Dublin, Ireland

https://doi.org/10.36505/ExLing-2006/01/0039/000039

Abstract

We present experiments on mono-lingual and cross-lingual articulatory feature recognition for English and German speech data. Our goal is to investigate to what extent it is possible to derive and reuse articulatory feature recognizers, whether particular features are better suited to this task. Finally whether this goal is practically achievable with the chosen machine learning technique and the selected set of speech signal descriptors.

Introduction

Earlier results on articulatory feature recognition suggest that Support Vector Machines perform better than Hidden Markov Models (HMM) (Macek et al., 2005). In frame based articulatory feature recognition the HMM approach does not benefit from its ability to model probabilistic dependencies, which makes it very useful in the task of speech recognition based on phone level descriptions (Kanokphara et al., 2006). This is due to limited dependency between adjacent frames in articulatory feature recognition.

We present experiments with support vector machines that use different types of kernels, a linear and two polynomial of different orders. All parameters were automatically extracted using the PRAAT analysis tool (Boersma and Weenink, 2006). Extraction of descriptive attributes was per-formed to obtain values of MFCCs with first and second order differences, for-mants with first order differences and bandwidths, distance between adjacent formants (F3−F2, F2−F1, F1−f0), and fundamental frequency. To extract these values we analyzed the data as sequences of 25ms windows with 10ms step length. We used the TIMIT corpus for English and the Phondat2 corpus for German. Both corpora are based on read speech.

In this paper, for TIMIT only dialect region 3 is used as the training set (102 speakers, 10 sentences each) while the whole core test set is used as the test set. The core test set, which is the abridged version of the complete test set, consists of 8 utterances from each of 24 speakers.

The Phondat2 corpus was split up into 11 speakers for training and 5 speakers for testing. Only sentences for which manual annotations were available were used (64 sentences per speaker).

ExLing 2006: Proceedings of 1st Tutorial and Research Workshop on Experimental Linguistics, 28-30 August 2006, Athens, Greece

Experiments

In this paper, SVMs with polynomial kernel were used for extraction of articulatory features from the speech signal. SVMs (Schoelkopf and Smola, 2002) learn separating hyperplanes to classify instances in the feature space that are mapped from the input space of the classified data. The mapping from input space to feature space is performed with application of a kernel. The dimension of the feature space is typically much higher than that of the original input space which allows for separability of the data.

The performance of the SVMs on feature recognition in German and English was compared. Performance of other methods on the same task was reported in (Kanokphara et al., 2006; Macek et al., 2005). The SVM classifiers were run with the SVMLight implementation (Joachims, 1999).

The speech signal was classified in a frame-by-frame manner, where every 10 ms for each frame of 25 ms length a set of descriptors was extracted. The non-speech parts, such as silence, are excluded in both feature sets in the training and evaluation. The performance of the articulatory feature recognition was evaluated for different orders of polynomial kernels of the SVMs up to order 3. The performance improved consistently with increasing order of the kernels. In this work we experimented with two sets of descriptors. The first consisted of 12 MFCC values, its first and second order differences, 5 formants (F1−F5) with first order differences and bandwidths, and pitch (f0). The second set of attributes extended the first one by distance in frequency between formants (F3−F2, F2−F1) and lowest formant and pitch (F1−f0).

The comparison of the two sets showed only negligible differences. The results presented in Table 1 are for the second set of descriptors used together with SVMs with polynomial kernel of order 3.

Results and discussion

To evaluate the performance of the recognizers we used two measures, namely the accuracy that gives overall performance regardless of the class distribution in the data, and the F_1-measure, which is the harmonic mean of the class dependent values of precision and recall. The F_1-measure gives a better picture of the actual performance of the recognizer for the relevant class. The F_1-measure is given priority for reasoning about the results.

In the mono-lingual setting, there are large differences in the performance, e.g. robust features such as [CONSONANTAL], [CONTINUANT], [SONORANT] and less well recognized features as [DORSAL], [LABIAL], [LATERAL]. In most cases a feature that is robustly recognized in the monolingual English setting is recognized robustly in the monolingual German setting, and nonrobust features are nonrobust in either of the two languages.

Table 1. Mono-lingual and cross-lingual recognition results.

	E-E		G-G		E-G		G-E	
Feature	Acc. in %	F_1-measure	Acc. in %	F_1-measure	Acc. in %	F_1-measure	Acc. in %	F_1-measure
−anterior	91.01	0.923	87.37	0.914	63.37	0.672	82.53	0.870
+anterior		0.892		0.764		0.586		0.734
−atr	92.75	0.957	79.18	0.843	56.39	0.661	77.20	0.861
+atr		0.758		0.690		0.390		0.374
−back	89.02	0.894	92.83	0.916	81.59	0.770	81.67	0.817
+back		0.886		0.937		0.847		0.816
−consonantal	89.10	0.885	94.24	0.931	88.36	0.851	85.11	0.853
+consonantal		0.896		0.951		0.904		0.849
−continuant	90.26	0.865	91.86	0.903	68.51	0.716	78.26	0.655
+continuant		0.924		0.930		0.646		0.841
−coronal	84.01	0.753	86.88	0.909	45.13	0.455	62.98	0.622
+coronal		0.882		0.764		0.447		0.637
−distributed	98.85	0.994	99.55	0.998	98.81	0.994	97.98	0.990
+distributed		0.710		0.792		0.265		0.160
−dorsal	93.02	0.962	91.62	0.953	88.55	0.937	88.59	0.938
+dorsal		0.574		0.583		0.348		0.228
−high	87.96	0.899	87.65	0.884	74.76	0.726	71.51	0.799
+high		0.850		0.868		0.766		0.511
−labial	84.56	0.906	91.55	0.952	82.86	0.897	81.26	0.888
+labial		0.561		0.667		0.495		0.425
−lateral	97.78	0.989	98.65	0.993	97.17	0.986	97.30	0.986
+lateral		0.368		0.133		0.003		0.004
−low	87.31	0.925	92.94	0.950	76.55	0.855	78.45	0.857
+low		0.581		0.880		0.385		0.565
−nasal	97.92	0.989	95.86	0.965	92.59	0.940	95.88	0.978
+nasal		0.810		0.949		0.904		0.652
−round	92.01	0.956	95.06	0.973	89.72	0.943	88.51	0.935
+round		0.603		0.718		0.468		0.491
−sonorant	96.69	0.978	96.35	0.972	93.23	0.947	85.50	0.909
+sonorant		0.937		0.948		0.905		0.642
−strident	92.99	0.704	97.19	0.984	59.78	0.699	77.31	0.439
+strident		0.960		0.904		0.393		0.858
−vocalic	93.12	0.936	94.29	0.952	84.84	0.883	88.10	0.880
+vocalic		0.926		0.929		0.783		0.882
−voiced	93.59	0.883	94.64	0.914	86.88	0.745	90.25	0.834
+voiced		0.956		0.961		0.912		0.931
−vot	90.01	0.884	93.84	0.966	64.36	0.776	41.67	0.561
+vot		0.912		0.656		0.128		0.130

In the cross-lingual settings, the most obvious pattern is that a recognizer that performs poorly in the mono-lingual setting will perform even worse in the cross-lingual setting. However, good performance in the mono-lingual setting is not general indicator for good cross-lingual performance. The features [CONSONANTAL], [VOCALIC], [VOICED] and [SONORANT] are the only ones that appear to be robust across all language combinations.

As can be seen in Table 1 accuracy can be a misleading measure of performance in the case of highly uneven class distribution in the data, an extreme example is the feature [LATERAL].

Although the original motivation for the articulatory features is their language independence, the comparison of performances on different features here suggests that some of the features are more language independent than others. However, studies such as that presented here do provide indications as to how features could be suited for the purposes of multilingual speech recognition.

Acknowledgements

This material is based upon works supported by the Science Foundation Ireland under grant No. 02/INI/II00. The opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of SFI.

References

Garofolo, JS, Lamel, LF, Fisher, WM, Fiscus, JG, Pallett, DS and Dahlgren, NL. 1993. DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CD-ROM. Technical Report NISTIR 4930, National Institute of Standards and Technology, Gaithersburg, MD, USA

Kanokphara, S, Macek, J, and Carson-Berndsen, J. 2006. Comparative Study: HMM & SVM for Automatic Articulatory Feature Extraction. In Proc. of the 19th Intern. Conf. on Industrial, Engineering & Other Applications of Applied Intelligent Systems, Annecy, France, June 2006. Springer Verlag.

Macek, J, Kanokphara, S and Geumann, A. 2005. Articulatory-acoustic Feature Recognition: Comparison of Machine Learning and HMM methods. In Proc. of the 10th Intern. Conf. on Speech and Computer SPECOM 2005, vol. 1, 99–103. University of Patras, Greece, 2005.

Joachims, T. 1999. Making large-scale SVM learning practical. In Schoelkopf, B, Burges, CJC and Smola, AJ (eds.), Advances in Kernel Methods—Support Vector Learning, 169–184, Cambridge, MA, MIT Press.

Schoelkopf, B and Smola, AJ. 2002. Learning with Kernels. Cambridge, MA, MIT Press.

Boersma, P and Weenink, D. 2006. Praat: doing phonetics by computer (Version 4.4.04) [Computer program]. Retrieved from http://www.praat.org/.